Model for RNA Replication and Mutation

1 Introduction

Here we establish a mathematical model to show how the RNA families amplified by Qβ replicase behave. The replication and mutation processes are modeled as a branching process. We study the mathematical expectation of mutated RNA products respect to time. Also, we use computer simulation to illustrate how the general mechanism works and how the RNA products could cover the libraries.

2 Notations

1. M_0: number of initial RNAs
2. μ: mutation rate, the probability that a single position would mutate during one time period
3. η: replication efficiency, the ration of RNAs that would be replicated during one time period
4. M_n: the number of generated RNA after n time period
5. A_n: the number of generated RNA that mutation does not happen in a given position after n time period
6. B_n: the number of generated RNA that mutation happens in a given position after n time period

3 Basic Assumption

We model the process of RNA replication and mutation as a branching process. In each discrete time period (or one cycle), every RNAs have a probability of η to

be replicated and every base in the new-generated RNAs has a probability of μ to mutate. Here we assume:

1. μ and η are constant.

2. Mutation are independent between different RNA families. For example, if there are 2 initial RNAs, X and Y, the mutation of RNAs generated from X will not affect the mutation of RNAs generated from Y. However, it is not difficult to notice that in a single RNA family, the mutation of different RNAs are actually not independent.

3. The probability of different mutation consequence for one position are equal. For example, the probability of a C mutates to A in one cycle is equal to the probability of a C mutates to G in one cycle, and equal to the probability of a U mutates to A in one cycle.

4 Computation and Analysis

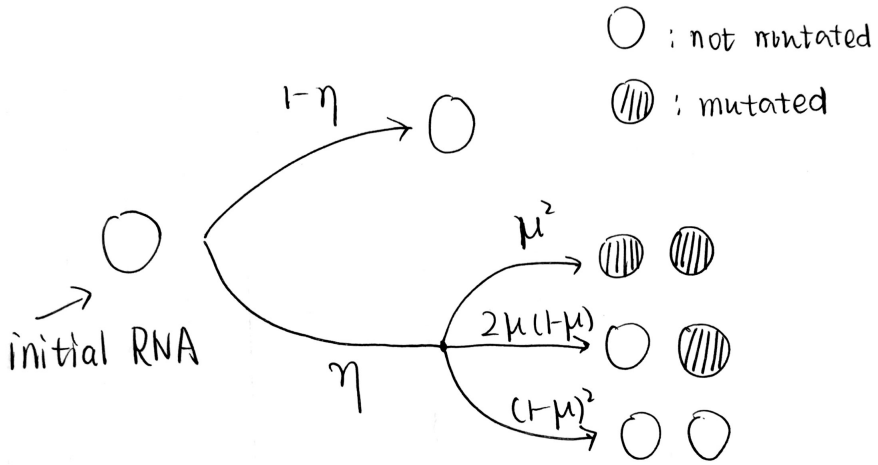
4.1 Expectation of number of RNA products which mutation happens in a given position

The first step is to analyze how single position in one RNA sequence mutates after several time periods pasts. Here we estimate the expectation of B_n, the number of generated RNA that mutation happens in the position after n time periods. To start with, we calculate the excepted number of M_n, the total number of RNA sequences after n time periods. By Assumption1, it is not difficult to establish a recurrence relationship for $\mathbb{E}(M_n)$:

$$\mathbb{E}(M_n) = (1+\eta)\mathbb{E}(M_{n-1}) = M_0(1+\eta)^n \tag{1}$$

It coincides with the exponential growth rate of RNA amplification by the Qβ replicase.

We could also establish a recurrence equation for A_n. To do this, we consider how the RNA behave in time period 1. For each initial RNA, there are four consequences:



By Assumption3 we have:

$$\begin{aligned}\mathbb{E}(A_n) =&(1-\eta)\mathbb{E}(A_{n-1}) + 2\eta(1-\mu)^2\mathbb{E}(A_{n-1})+ \\ &2\eta\mu(1-\mu)\left(\mathbb{E}(A_{n-1}) + \frac{\mathbb{E}(B_{n-1})}{3}\right) + 2\eta\mu^2\frac{\mathbb{E}(B_{n-1})}{3}\end{aligned} \tag{2}$$

and we must have:

$$\mathbb{E}(A_n) + \mathbb{E}(B_n) = M_0(1+\eta)^n \tag{3}$$

Now we are able to express $\mathbb{E}(A_n)$ in explicit formula. Solve equation (2) and (3), we have

$$\mathbb{E}(A_n) = \frac{M_0}{2}\left((1+\eta)^n + (1+\eta-4\eta\mu)^n\right) \tag{4}$$

and

$$\mathbb{E}(B_n) = \frac{M_0}{2}\left((1+\eta)^n - (1+\eta-4\eta\mu)^n\right) \tag{5}$$

The ratio of RNA products that has mutated in this position R would be:

$$R = \frac{1}{2}\left(1 - \left(1 - \frac{4\eta\mu}{1+\mu}\right)^n\right) \tag{6}$$

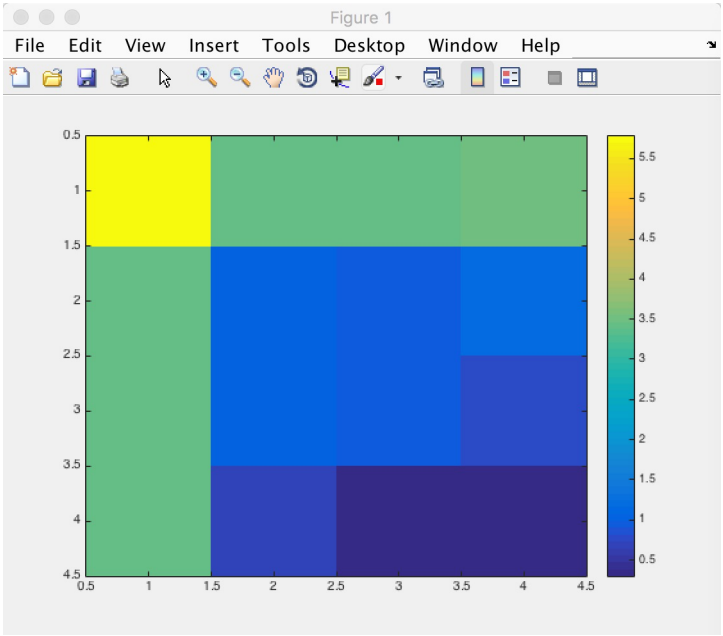
R will tend to $\frac{1}{2}$ as n tends to big enough, although in practice we can imagine that the convergence speed would be quite slow. Nevertheless, it is enough to point out that **for one position or one base**, as time passes by, the mutated RNA would take a significant proportion. Or in other words, for the whole sequence, for every positions we could find a certain proportion of RNAs in which mutation happened in this position after several time period.

4.2 Computer simulations for 2 bases

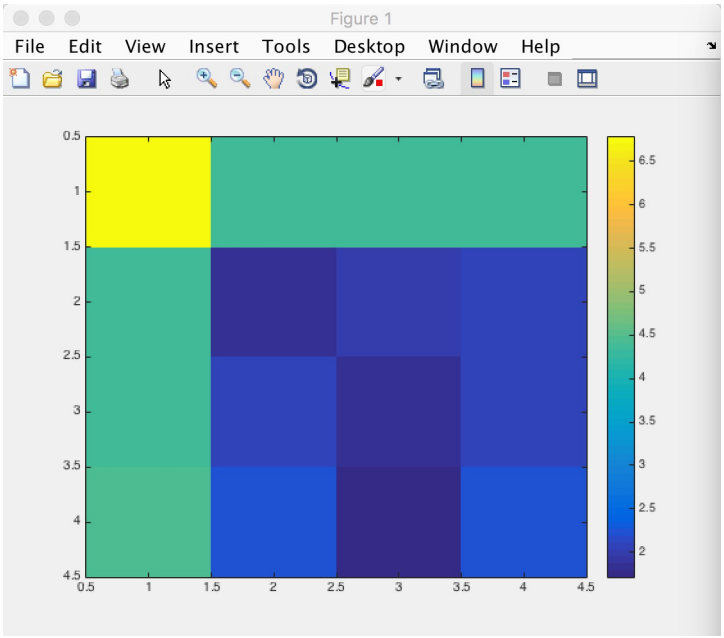
We now consider two bases, p_1 and p_2, in RNA sequence. As we mentioned in Assumption3 that the mutation are not totally independent, so we cannot directly use our previous conclusion. Instead, we use a computer simulation first. We construct a 4*4 matrix $A(n) = [a_{ij}(n)]$:

$$a_{ij}(n) = k, \text{if in time } n, p_1 \text{ is } i \text{ and } p_2 \text{ is } j \tag{7}$$

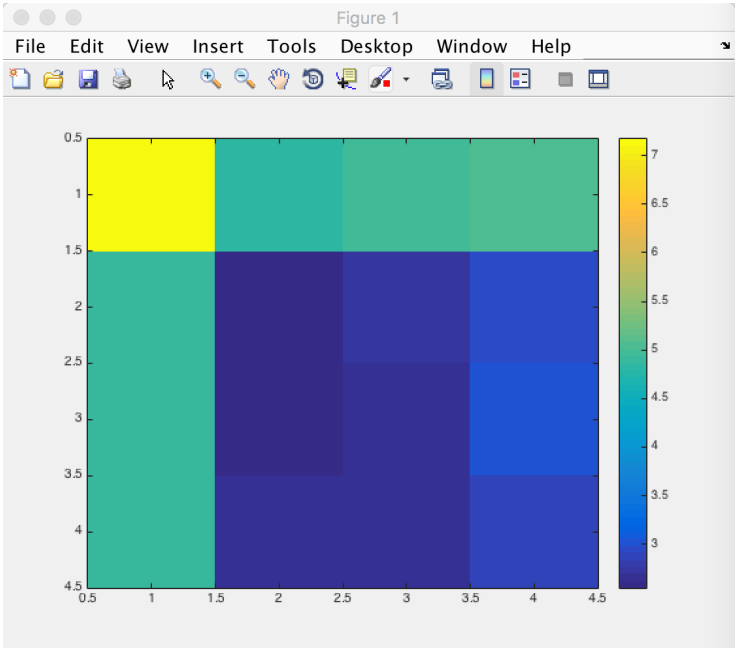
where 1,2,3,4 denotes four nucleotide, A, U, C, G, respectively. We set initially, $p_1 = p_2 = A$, or $a_{11} = M_0$. We use number of μ by Bradwell et al. (2013) as 0.0013 and set $\eta = 0.9$, M_0=1000, 10000 and $n = 10, 15$. We also compute the $\log_{10} a_{ij}$ to illustrate the order of magnitude. The MATLAB code for simulation is provided later. Here we show the results:



1.M_0=1000, n=10 The graph (and the following graphs) shows the order of magnitude (or $\log_1 0$) of $A(8)$. Most RNAs remain unchanged. Approximately tens of thousand RNAs changed one position, and others (dozens of) changed both position. All possible consequences observed.



2.M_0=10000, n=10 The order of magnitude for each term in the matrix increased by one (that is, 10 times) compared to when $M_0 = 1000$. For example, the number of RNAs that changed both positions is 50-150 for each possible type.



3.M_0=1000, n=15 The order of magnitude for each term in the matrix increased even more significantly. For example, the number of RNAs that changed both positions is 380-1000 for each possible type.

Also, the number of A_n are well consistent with our expectation. For example, for M_0=1000, n=10, $\mathbb{E}(A_n)$ by formula(4) is 605638 and A_n is 599252 in our simulation.

4.3 Conclusion

We have shown that for a specific position, the mutation are relatively easy to reach and the proportion of mutated RNA could be very significant (the green part in the simulation graphs). For two bases, if we want to reach a mutation for both two bases, it is still feasible to increase M_0, the number of initial RNAs and n, the reaction time (more effective) to reach a substantial number of expected mutation. Although theoretically the RNAs could mutate to any given type as n goes to big enough, there are certain problems like the break up of exponential growth rate (due to the consumption of resources) and the time or initial number of RNAs to reach a given result would be too tremendous. Nevertheless, to cover any types of relatively smaller change (like 1, 2 bases as we shown, or 3 bases which is reasonable to assume), it is available to use $Q\beta$ replicase to reach an expected result.

5 MATLAB code

```
seqNum=zeros(4,4);
seqNum(1,1)=5000; %number of initial RNAs
mutR=0.001; %muatation rate
repR=0.9; %replication rate
cycle=10; %number of cycles
temp=seqNum;
for n=1:cycle
    for i=1:4
        for j=1:4
            if seqNum(i,j)>0
                for k=1:seqNum(i,j)
                    if rand()<repR
                        temp(i,j)=temp(i,j)-1;
                        if rand()<mutR
```

```
                        t = [1:(i-1),(i+1):4];
                        first = t(randperm(3,1));
                    else
                        first = i;
                    end
                    if rand()<mutR
                        t =[1:(j-1),(j+1):4];
                        second = t(randperm(3,1));
                    else
                        second = j;
                    end
                    temp(first,second)=temp(first,second)+1;
                    if rand()<mutR
                        t = [1:(i-1),(i+1):4];
                        first = t(randperm(3,1));
                    else
                        first = i;
                    end
                    if rand()<mutR
                        t = [1:(j-1),(j+1):4];
                        second = t(randperm(3,1));
                    else
                        second = j;
                    end
                    temp(first,second)=temp(first,second)+1;
                end
            end
        end
    end
end
seqNum=temp;
end
disp(seqNum)
imagesc(log10(seqNum))
colorbar
```

6 reference

Bradwell, K., Combe, M., Domingo-Calap, P. and Sanjuan, R. (2013). Correlation Between Mutation Rate and Genome Size in Riboviruses: Mutation Rate of Bacteriophage Q. Genetics, 195(1), pp.243-251.

Kimmel, M. and Axelrod, D. (2015). Branching processes in biology.